SHORTER ARTICLES AND DISCUSSION

COMPUTING CORRELATION IN CASES WHERE SYMMETRICAL TABLES ARE COMMONLY USED

In studying the assortative mating of *Paramecium* I have found occasion to compute the correlation in many cases for which double or symmetrical tables are commonly employed. I have found that in such cases the use of symmetrical tables is quite unnecessary and the computations can be performed with much less labor without them. It will, therefore, be worth while to show how the use of symmetrical tables can be avoided.

When the two objects to be compared are alike, as when the two members, A and B, of conjugating pairs are examined, evidently either A or B might be entered in either the horizontal rows or the vertical columns of the correlation table. In such cases, the mean computed from the rows, and that computed from the columns are likely to differ, depending on which individuals were entered in the rows, which in the columns. If, for example, the larger individual is always entered in the vertical columns, the smaller in the horizontal rows, as in Table II, then the means and standard deviations of the two sets will differ much. As a result the coefficient of correlation computed in the usual way will show varying values, depending on how the pairs are entered in the table. From the collection shown in Table II we can by varying the method of entering the pairs get coefficients of correlation varying from 0.132 to 0.523.

Under such conditions Pearson (1901), Pearl (1907) and others enter each pair twice, once in the rows, once in the columns. This gives a "symmetrical" table, in which the sums of either the rows or the columns include all the individuals. This method is theoretically correct, since each individual functions both as "principal" and as "mate"; the coefficient of correlation computed from such symmetrical tables is the correct one. But such symmetrical tables are cumbersome and involve much labor. Pearl (1907) gives a formula by which the same coefficient can be obtained without making symmetrical tables, by computations involving the two means and standard

deviations and the coefficient of correlation found in the usual way.

But it is possible to find the correct coefficient of correlation from ordinary tables, and with much less labor than by either the use of symmetrical tables or by the method given by Pearl. To see how this can be done, it is well to examine a symmetrical table prepared for computation of the coefficient of correlation, such as is given in Table I. Here the large figures give the frequencies, while the subscripts in smaller type give the products of the deviations from the approximate mean (37). There are two main points to be considered: (1) How the quantity

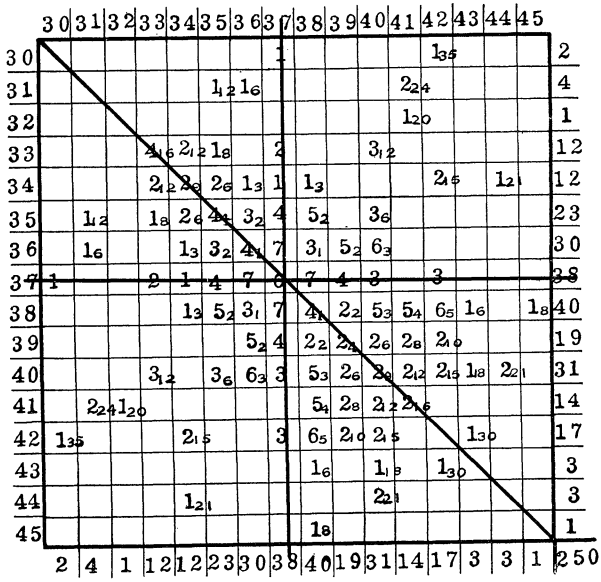

TABLE I. SYMMETRICAL CORRELATION TABLE FOR THE LENGTHS OF 125 PAIRS OF *Paramecium aurelia;* each individual entered twice, once in the vertical columns, once in the horizontal rows. (Unit of measurement, 4 microns.)

$S(xy)$ is to be correctly obtained; (2) how the mean and standard deviation are to be correctly obtained.

1. With regard to the first point, it will be observed that such a table is divisible by a diagonal passing from the upper left-hand corner to the lower right-hand corner into two halves which are in all respects duplicates as regards both frequencies and deviation products. (The frequencies through which the diagonal line passes are to be divided evenly between the two halves.) It is evident, therefore, that if we use only one of these halves

of the table in getting the sum $S(xy)$ we shall get just one half the sum we should get by using the whole table; the sum for the whole table would therefore be obtained simply by doubling this half-sum. Now, if in place of making a symmetrical table we enter always the larger member of each pair in the vertical columns, the smaller in the horizontal rows, we shall get a table that is precisely one of these duplicate halves of the symmetrical table; this will be seen by comparing Tables I and II. The quantity $S(xy)$ from such a table will then be just half that from the symmetrical table; it may then be doubled, and the further computation will be identical with that for the symmet-

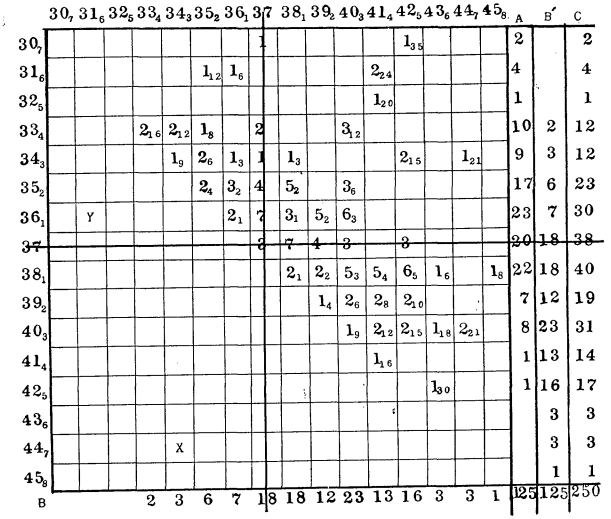

TABLE II. THE SAME TABLE SHOWN IN TABLE I, SAVE THAT EACH INDIVIDUAL IS ENTERED BUT ONCE—the larger member of the pair in the vertical column, the smaller in the horizontal row.

rical tables. Or (as we shall see) this half sum, which forms the dividend in obtaining the coefficient of correlation, may be divided by a number half as great as in the symmetrical tables, giving the same result.

It will further be seen that if in place of entering all pairs in the same way—the larger members in the columns, the smaller in the rows—we enter some or all of them differently, this will make no difference in the result. If in Table II, for example, the pair showing measurements 44 by 34 were entered in the reverse way, it would fall, no longer in the right upper quad-

rant, but in the left lower quadrant, at the point marked *X*. Here, as examination will show, it would receive the same subscript that it has now, and would count as negative, exactly as it now does. Again, suppose the pair 36 by 31 were similarly transposed; it would still fall in the left upper quadrant, at the point marked *Y*, where it would receive the same subscript as at present and count as positive, just as at present. And so of all other cases; the value of a pair is not altered in any way by changes in the way it is entered in the table. In making the table, therefore, the pairs may be entered only once and quite at random, or in any way that is convenient.

2. With regard to the mean and standard deviation, the apparent advantage of symmetrical tables is that they give us the actual mean of all the individuals; it is to this mean that our correlation must refer. But this actual mean can readily be obtained from the tables in which each pair is entered but once, in any way that happens to be convenient. It is merely necessary to add together the sums of the rows and of the columns of the table. Thus in Table II the number of individuals having the length 35 is not 17 (sum from the row beginning with 35), nor 6 (sum from the column headed 35), but 23 (sum from both the row and the column) and so for all other classes. It will be well to illustrate by an example certain of the steps in the computation. Table II shows a correlation table of single entry, as prepared for computation of the coefficients of correlation and other constants.

After finding the sums of the rows (given in column A at the right) and of the columns (given in B, underneath), we place the latter sums (B) by the side of A, in the proper places (as at B′), then add the two sets, giving the sums shown in the column C at the right. These are the same sums that we should get from a symmetrical table; adding these we get the total number of individuals (250 in Table II). Now from this column C we find the approximate mean in the usual way; it lies in this case at the length 37 (with 38 individuals). Through the column and the row headed 37 we therefore draw the lines serving as axes of reference in finding the correlation. We now find the correlation in the usual way. In so doing (1) we make use always of the sums in the column C in finding mean, standard deviation, etc. (2) We use for both horizontal and vertical axes of reference in computing the correlation in all cases

a row and column with the same heading (37 in this case). (3) We employ the ordinary frequencies in the body of the table in getting the sum of the deviations of (xy) for use in the formula for the coefficient of correlation, just as in ordinary correlation tables. The computation of the coefficient is of course (as in the case of symmetrical tables) considerably simpler than in the usual case, since we have but one standard deviation and one quantity d to deal with.

Only one other point in the computation is peculiar, requiring careful observance. If we let n signify the number of pairs and N the number of individuals (so that $N = 2n$), then in finding the mean, standard deviation, and coefficient of variation, we use N (just as in symmetrical tables), so that the formula for the standard deviation is

$$\sigma = \sqrt{\frac{S(x^2)}{N} - d^2} - .083.$$

But in getting the coefficient of correlation, the sum $S(xy)$ which we get from our unsymmetrical table is just half what we should get from a symmetrical table (as we have already seen). Therefore, to make the computations identical with those for symmetrical tables, we must either double this sum in the formula for the coefficient of correlation, or what is simpler, in place of doubling this sum we may halve the number by which we divide this sum, that is, we may use n in place of N. Thus the formula for the coefficient of correlation becomes by this method

$$r = \left(\frac{S(xy)}{n} - d^2\right) \times \frac{1}{\sigma^2}.$$

This method lends itself readily to the valuable procedure recently described by Harris (1910) for finding the coefficient of correlation, the only point requiring careful attention being the fact that in finding the standard deviation we must use N (number of individuals), while in the formula for the coefficient of correlation we must use n (number of pairs). The present plan is likewise well adapted for finding the coefficient of correlation by the "difference method" (see Harris, 1909).

If the method we have described is used, the pairs are entered in the table but once, in any way that is convenient; the correlation computed will always be the same, and identical with that from symmetrical tables. It avoids the cumbersome and labo-

rious symmetrical table; at the same time it involves much less labor than the method given by Pearl. When there are many tables to be computed, the amount of drudgery it saves is great.

PAPERS CITED

1909. Harris, J. A. A Short Method of Calculating the Coefficient of Correlation in the Case of Integral Variates. *Biometrika*, 7, 215–218.

1910. Harris, J. A. The Arithmetic of the Product Moment Method of Calculating the Coefficient of Correlation. *Amer. Naturalist*, 44, 693–699.

1907. Pearl, R. A Biometrical Study of Conjugation in Paramecium. *Biometrika*, 5, 213–297.

1901. Pearson, K. Mathematical Contributions to the Theory of Evolution—IX. On the Principle of Homotyposis and its Relation to Heredity, to the Variability of the Individual and to that of the Race. *Philos. Trans.*, A, 197, 285–379.

THE JOHNS HOPKINS UNIVERSITY.

H. S. JENNINGS.